PPD pipeline

Part1: generate "degenerated" sequences

Part2: trimming and paralog detection

Gene: "*degenerated*" MSA

Intron 1	Exon	Intron 2

Sp1 — G — W — S — Y — K — R
Sp2 — G — W
Sp3 — G — W
Sp4 — T — T
Sp5 — G — W
Sp6 — G — W
Sp7 — T — W
Sp8 — T — T — TACATGGTTCTA
Sp9 — T — W
OG — KC T — T
Target gene — T

Polymorphic site 1   Polymorphic site 2   Hypervariable sites

PPD Trimming Steps
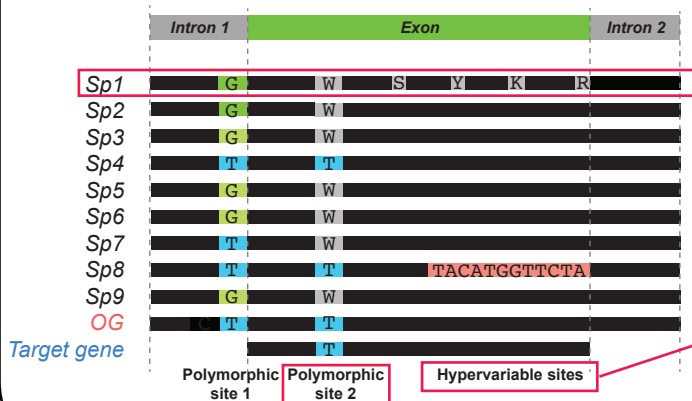
S1: resort gene files

S2: filter sequences with high percentage of heterozygous sites (default 0.05%).

S3: add target references to gene files.

S4: obtain MSA using MAFFT.

S5: remove the target references.

S6: remove the sites with over 50% gaps (default) using Trimal.

S7: remove hypervariable sites using sliding window method.

Paralogs Detection

S8: identify putative paralogs using shared heterozygosity in over 50% taxa.